MasonTigers@LT-EDI-2024: An Ensemble Approach Towards Detecting Homophobia and Transphobia in Social Media Comments

Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan,
Al Nahian Bin Emran
George Mason University, USA
{dgoswam, spuspo, mraihan2, abinemra}@gmu.edu

Abstract

In this paper, we describe our approaches and results for Task 2 of the LT-EDI 2024[1] Workshop, aimed at detecting homophobia and/or transphobia across ten languages. Our methodologies include monolingual transformers and ensemble methods, capitalizing on the strengths of each to enhance the performance of the models. The ensemble models worked well, placing our team, *MasonTigers*, in the top five for eight of the ten languages, as measured by the macro F1 score. Our work emphasizes the efficacy of ensemble methods in multilingual scenarios, addressing the complexities of language-specific tasks.

1 Introduction

In this current era dominated by social media platforms, people heavily rely on online content for communication, learning, knowledge-sharing, and staying abreast of new technologies. The comment sections, intended for constructive feedback, unfortunately, sometimes become grounds for hate speech, offensive comments, and discrimination, including targeting a specific community. Such behaviors cause trauma, fear, anxiety, depressive symptoms and discomforts among LGBTQ+ individuals (Poteat et al., 2014; Ventriglio et al., 2021), hindering them from freely expressing their thoughts and feedback.

To ensure the safety and comfort of users on online platforms, it becomes imperative to identify and address hate speech and offensive comments. Although there are existing policies aimed at protecting communities from such misconduct, violations may lead to the removal of offending comments [2] [3]. However, the identification process necessitates the application of NLP and AI techniques due to the diverse nature of hate speech, which can manifest in both direct and passive forms. Surprisingly, there is a higher focus on researching this topic in English and other high-resource languages like hindi, with ample resources. For instance, couple of shared tasks have been organized previously e.g. Chakravarthi et al. (2022b), Chakravarthi et al. (2022c), Chakravarthi et al. (2023), accompanied by substantial datasets e.g. Vásquez et al. (2023), Chakravarthi et al. (2021). However, there has been a lack when it comes to identifying hate speech in low-resource and under-resource languages.

This shared task Chakravarthi et al. (2024) aims to identify hate speech contents, specifically homophobia, transphobia, and non-anti-LGBT+ sentiments, directed at LGBTQ+ individuals in 10 different languages, including 7 low-resource languages. To tackle the linguistic diversity, we conduct separate experiments for each language, leveraging different transformer-based models with proficiency in distinct languages. Notably, for Tulu, an under-resourced language, we employ a prompting approach. Alongside these experiments, various techniques are explored, and the most effective one during the evaluation phase is implemented in the test phase for comprehensive validation.

2 Related Works

As smart devices, mobile apps, and social media platforms become more widely utilized, there are also more negative effects linked to them, such as cyberbullying, hurtful comments, and rumors have increased. These platforms have also become a space for negative behaviors like sexism, homophobia (Diefendorf and Bridges (2020)), misogyny (Mulki and Ghanem (2021)), racism (Larimore et al. (2021)) and transphobia (Giametta and Havkin (2021)). Internet trolling, where people say mean things online, has become a global problem. To deal with this, researchers are looking

---

into automated methods since checking every message manually is impossible. In this section, we provide a brief summary of the research attempts that focused on identifying homophobia, transphobia, and non-anti-LGBT+ content from YouTube comments.

To motivate researchers to tackle the issue of identifying toxic language directed at the LGBTQ community, in the last few years, several shared tasks have been released. In one such shared task related to LT-EDI-2022 (Chakravarthi et al.), researchers submitted systems to deal with homophobic and transphobic comments. Chakravarthi et al. (2022b) gives an overview of the models submitted. Three subtasks for the languages Tamil, English, and Tamil-English (code-mixed) were the emphasis of this shared task. Apart from this, several studies have been conducted. The author of Karayigit et al. (2022) used the M-BERT model, which shows that it is capable of accurately identifying homophobic or other abusive language in Turkish social media comments. Similarly, another shared task related to this topic has been organized showing XLM-R performing best with spatio-temporal data in 5 different languages (Wong et al., 2023). Vásquez et al. (2023) presents a mexican-spanish annotated corpus along in which beto-cased (Spanish BERT) outperforms the other models.

For several text classification tasks, transformer-based ensemble approaches perform very well, like the works by Goswami et al. (2023b); Raihan et al. (2023b). Also, prompting Large Language Models like GPT3.5 (OpenAI, 2023) is another popular approach in recent classification tasks (Raihan et al., 2023c) for the past year.

Efforts to identify homophobic and transphobic comments have primarily focused on a maximum of five languages to date. However, in this shared task, a total of 10 languages have been chosen, and ongoing efforts now include Telugu, Tulu, and Marathi.

3 Datasets

The dataset provided for the shared task contains 10 languages - Tamil, English, Malayalam, Marathi, Spanish, Hindi, Telugu, Kannada, Gujarati, and Tulu. It is compiled using five separate research works. The previous iteration of the workshop (Chakravarthi et al., 2023) includes Tamil, English and Spanish, Hindi, Malayalam languages, and the earlier version Kumaresan et al. (2024) includes Tamil, English, and Tamil-English (Code-Mixed) languages.

The work by Kumaresan et al. (2023) builds a dataset for Malayalam and Hindi languages from social media comments. Another dataset by Chakravarthi (2023) focuses on YouTube comments. One data augmentation approach is adopted by Chakravarthi et al. (2022a).

The current dataset combines all these works to build a comprehensive dataset for the task of homophobia and/or transphobia detection in 10 languages Chakravarthi et al. (2024).

The detailed dataset demonstration and label-wise data percentage for all the languages are available in Table 1.

4 Experiments

Among the 10 languages (except Tulu) we use XLM-R, m-BERT and language specific BERTs: RoBERTa (Liu et al., 2019) , HindiBERT (Nick Doiron, 2023), TamilBERT (Joshi, 2023), Malayalam-BERT (Joshi, 2023), MarathiBERT (Joshi, 2022), SpanishBERT (Cañete et al., 2022), TeluguBERT (Joshi, 2023), KannadaBERT (Joshi, 2023), GujaratiBERT (Joshi, 2023) for Hindi, Tamil, Malayalam, Marathi, Spanish, Telugu, Kannada, Gujarati respectively. While using MarathiBERT and HindiBERT, we pad the sentence length upto 512 tokens, because of the limitation of the aforementioned BERTs. Training parameters are mostly kept the same across all models, mentioned in Table 3.

After that we perform weighted ensemble ap-

Role: "You are a helpful AI assistant. You are given the task of detecting homophobia and transphobia in a given text.

Definition: Homophobia and transphobia detection is the process of identifying expressions of hatred or discrimination against LGBTQ+ individuals in communication.'.

Examples: An example of Homophobic/Transphobic comment: **<Example1>**. An example of Non-Homophobic/Transphobic comment: **<Example2>**'.

Task: Generate the label [YES/NO] for this **"text"** in the following format: *<label> Your_Predicted_Label <\label>*. Thanks."

Figure 1: Sample GPT-3.5 prompt for few shot learning [*Used for the* **Tulu** *Dataset*].

Tamil				English			
Labels	Train	Dev	Test	Labels	Train	Dev	Test
Non-anti-LGBT+ content	77.53	76.13	76.11	Non-anti-LGBT+ content	94.12	94.45	94.04
Homophobia	17.02	17.72	18.25	Homophobia	5.66	5.30	5.56
Transphobia	5.45	6.15	5.64	Transphobia	0.22	0.25	0.40
Malayalam				**Marathi**			
Labels	Train	Dev	Test	Labels	Train	Dev	Test
Non-anti-LGBT+ content	79.25	77.25	77.83	None of the categories	73.49	72.13	75.87
Homophobia	15.29	16.24	16.17	Homophobia	15.74	17.20	14.93
Transphobia	5.46	6.51	6.00	Transphobia	10.77	10.67	9.20
Spanish				**Hindi**			
Labels	Train	Dev	Test	Labels	Train	Dev	Test
None	58.34	51.82	50.00	Non-anti-LGBT+ content	94.65	95.31	95.95
Transphobic	20.83	24.09	25.00	Transphobia	3.59	4.06	3.12
Homophobic	20.83	24.09	25.00	Homophobia	1.76	0.63	0.93
Telugu				**Kannada**			
Labels	Train	Dev	Test	Labels	Train	Dev	Test
None of the categories	38.63	38.51	38.37	None of the categories	44.35	44.27	44.11
Homophobia	32.12	31.18	32.18	Homophobia	28.17	28.61	28.11
Transphobia	29.25	30.31	29.45	Transphobia	27.48	27.12	27.78
Gujarati				**Tulu**			
Labels	Train	Dev	Test	Labels	Train	Dev	Test
None of the categories	47.39	45.29	45.63	NON H/T	74.25		82.32
Homophobia	27.93	28.62	29.31	H/T	25.75		17.68
Transphobia	24.68	26.09	25.06				

Table 1: Label-wise Data Percentage for Different Languages

proach of the above models (XLM-R, m-BERT, language specific BERTs) and use the macro F1 scores of the models on the dev data as the weight along with the confidence score to get the ensemble macro F-1 score of the test phase. After the test labels get published, we use the F1 score in the rank list as the weight along with the corresponding confidence score for the additional experiment of weighted ensemble approach.

As Tulu is very close to Kannada and Tulu doesn't have any language specific fine-tuned model, we used KannadaBERT on Tulu. In South Karnataka, individuals who speak Tulu are typically fluent in both Tulu and Kannada. Due to the long-standing interaction between Tulu and Kannada, it can be anticipated that codeswitching between these two languages is a probable outcome (Shetty, 2003). Moreover, we implement few shot learning using GPT3.5 for the Tulu dataset (see Figure 1). Such prompting is very widely used in recent works on text classification (Raihan et al., 2023a; Goswami et al., 2023a). We got the same result as the ensemble approach in the few shot prompting technique. We specifically used 8 - shot

prompting for Tulu language. This process is inspired by Wei et al. (2022).

We improve the macro F1 in 4 out of 10 cases in this additional experiments and rest of the 6 cases we get the same macro F1 as the test phase.

5 Results

We employ an ensemble-based methodology for the tasks, since in text classification tasks this can further improve the results. for For all the ensemble approaches, we use XLM-R, mBERT, and a BERT-based model fine-tuned for that specific langauge as we mentioned in the previous section.

In the testing phase, we ensemble the confidence score of the three models and then calculate the weighted average. The weight in this context is the macro F1 score of the corresponding models on dev data.

For Tulu language, we use GPT 3.5 for few shot prompting. We use few instances of each of the two labels with specific prompt and and then get the test labels predicted by GPT 3.5.

After the end of testing phase when the test labels get published, we further run the ensemble

Tamil (Rank 5)				English (Rank 10)		
Models	Dev F1	Test F1		Models	Dev F1	Test F1
XLM-R	0.49	0.49		XLM-R	0.32	0.32
mBERT	0.62	0.67		mBERT	0.32	0.32
TamilBERT	0.51	0.53		roBERTa	0.32	0.32
Wt. (Dev F1) Ensemble		0.51		Wt. (Dev F1) Ensemble		0.32
Wt. (Test F1) Ensemble		**0.52**		Wt. (Test F1) Ensemble		0.32

Malayalam (Rank 9)				Marathi (Rank 4)		
Models	Dev F1	Test F1		Models	Dev F1	Test F1
XLM-R	0.51	0.51		XLM-R	0.49	0.49
mBERT	0.54	0.55		mBERT	0.46	0.41
MalayalamBERT	0.52	0.52		MarathiBERT	0.41	0.44
Wt. (Dev F1) Ensemble		0.51		Wt. (Dev F1) Ensemble		0.44
Wt. (Test F1) Ensemble		**0.52**		Wt. (Test F1) Ensemble		**0.45**

Spanish (Rank 3)				Hindi (Rank 4)		
Models	Dev F1	Test F1		Models	Dev F1	Test F1
XLM-R	0.75	0.55		XLM-R	0.33	0.33
mBERT	0.79	0.55		mBERT	0.33	0.33
SpanishBERT	0.81	0.50		HindiBERT	0.33	0.33
Wt. (Dev F1) Ensemble		0.50		Wt. (Dev F1) Ensemble		0.33
Wt. (Test F1) Ensemble		**0.54**		Wt. (Test F1) Ensemble		0.33

Telugu (Rank 1)				Kannada (Rank 3)		
Models	Dev F1	Test F1		Models	Dev F1	Test F1
XLM-R	0.97	0.97		XLM-R	0.95	0.95
mBERT	0.96	0.95		mBERT	0.93	0.94
TeluguBERT	0.97	0.97		KannadaBERT	0.95	0.95
Wt. (Dev F1) Ensemble		0.97		Wt. (Dev F1) Ensemble		0.95
Wt. (Test F1) Ensemble		0.97		Wt. (Test F1) Ensemble		0.95

Gujarati (Rank 5)				Tulu (Rank 3)		
Models	Dev F1	Test F1		Models	Dev F1	Test F1
XLM-R	0.94	0.94		GPT 3.5		0.45
mBERT	0.95	0.95		XLM-R	0.42	0.42
GujaratiBERT	0.93	0.93		mBERT	0.42	0.45
Wt. (Dev F1) Ensemble		0.94		KannadaBERT	0.42	0.45
Wt. (Test F1) Ensemble		0.94		Wt. (Test F1) Ensemble		0.45

Table 2: Combined Results for Various Languages

Parameter	Value
Learning Rate	$1e-5$
Train Batch Size	8
Test Batch Size	8
Epochs	5

Table 3: Training Configuration Parameters

systems again for all the languages. For this case, the weight for the models is the macro F1 score of the corresponding models on test data. For Tulu language, we also peform this ensemble approach

with XLM-R, mBERT and KannadaBERT. Though none of these models have Tulu language in their corpora but all these models were pretrained on Kannada which is really close to Tulu. We achieve a macro F1 score using this ensemble approach which is as same as the few shot prompting.

Using this approach, we achieve the rank 1 (one) on Telugu, Rank 3 (three) on Spanish, Kannada and Tulu, rank 4 (four) on Marathi and Hindi, rank 5 (five) on Tamil and Gujarati, rank 9 (nine) on Malayalam and rank 10 (ten) on English language. The detailed experimental results of the models are available in Table 2.

6 Error Analysis

To thoroughly investigate the results and the models' performance on specific datasets, we find that though the accuracy of the models on all the datasets are very good but the macro F1 score is really low in some cases. From the table in Table 1, it is clearly visible that English and Hindi dataset is very imbalanced. They have a very few Homophobia and Transphobia label. From the confusion matrices in Appendix A, we can see all the instances of Non-anti-LGBT+ content label are correctly predicted by the models but models' fail to be well-trained on the other two labels. Thus the other two labels get mis-classified which leads to a marco F1 score around 0.33 for these two languages. For Tamil, Malayalam, Marathi, Spanish and Tulu - the data ratio is comparatively balanced which leads to F1 score in the range 0.45 - 0.54. Telugu, Kannada and Gujarati datasets are evenly label-wise evenly balanced which get reflected with highest marco F1 score in the range 0.94 - 0.97. For the imbalanced datasets, widely used techniques like data augmentation, back translation can be proven helpful which can be potential future research scope in this domain. Detailed error analysis for the languages is given below:

- **Tamil** The Homophobia instances are partially correct but all the instances of Transphobia are misclassified. On the other hand, the Non-anti-LGBTQ + content instances are almost perfectly classified in Tamil.

- **English** The models perform well in identifying only Non-anti-LGBT+ content while they completely fail to detect Homophobia, and Transphobia in English.

- **Malayalam** For Malayalam, the models almost perfectly detect Non-anti-LGBTQ + Content, partially detect Homophobia instances but as before it completely misclassified the Transphobia.

- **Marathi** The Homophobia and None of the categories instances are partially correct in case of Marathi. However, the transphobia content instances are mostly misclassified.

- **Spanish** Homophobic, None and Transphobic instances are partially correct in Spanish.

- **Hindi** Only Non-anti-LGBTQ + content instances are perfectly classified but all the in-

stances of Homophobia and Transphobia are misclassified in Hindi.

- **Telugu** Homophobia, None of the categories and Transphobia instances are almost perfectly correct in Telugu.

- **Kannada** The model predictions are almost correct in Kannada for instances of Homophobia, None of the categories and Transphobia.

- **Gujarati** Our model can almost perfectly detect for Homophobia, None of the categories for Gujarati Transphobia instances.

- **Tulu** Though NON H/T instances are perfectly classified, but all the instances of H/T are misclassified in Tulu.

7 Conclusion

The task of detecting abusive speech targeting sexual and gender minorities has become increasingly important given the rise of social media and its potential to propagate harmful stereotypes that further marginalize vulnerable populations. This paper presents our efforts to address online transphobia and homophobia in multilingual contexts, which remains an under-studied area in abusive language detection. We employ an ensemble approach combining multiple individual models to identify abusive speech across datasets in ten languages.

Our findings demonstrate that while no individual system consistently achieves superior performance across all data, monolingual language-specific BERT models fine-tuned on our abusive speech data unsurprisingly emerge as strong approaches for this classification problem. However, our ensemble framework leveraging voting across multiple BERT variants along with other models surpasses any individual system, indicating the value of model diversity even when one base technique manifests strengths. We hypothesize that the inconsistencies across models and languages result partly from imbalanced, sparse instances of actual abusive samples in our data. Hence, future work should prioritize constructing larger, more balanced benchmark datasets for abusive language detection encompassing underrepresented identities. Nonetheless, this research presents a starting point for identifying multilingual, multidirectional abuse in online spaces through ensemble natural language systems.

Limitations

The monolingual BERT models underperformed due to insufficient data volume and class imbalance in our existing abusive speech corpora. Skewed label distributions with far fewer minority abuse cases than benign texts make learning discriminative patterns difficult. Our ensemble framework mitigated these weaknesses but still suffered performance issues on minority samples. Constructing larger, more balanced datasets remains imperative yet challenges exist regarding sensitivity of human annotations for this problem space. Nonetheless, enhancing model robustness on sparse abusive instances should be prioritized. While augmenting through back-translation and generation could help, this risks polluting training data if new variants stray from actual phenomenology. Systems producing false positives that ascribe nonexistent abuse contribute harm. Progress on mitigating imbalance without these downsides is incremental. Our experiments manifested datasets with endemic skew away from minority classes. Researchers must remain cognizant that efforts to populate abusive classes risk drifting from reality.

Ethics Statement

Adhering to the ACL Ethics Policy, this study seeks to responsibly progress online safety through benign content filtering technology. However, safeguards against misuse for censorship/monitoring remain imperative. While the supplied dataset was anonymized to protect privacy, carefully compiled public benchmarks avoiding marginalization must drive future progress. Flawed training data has propagated harm before; our experiments mitigated this but work continues. Guiding principles of beneficence and nonmaleficence should steer research on automating content classification with real-world impacts, including on complex offensive speech. Assessing for unintended consequences and awareness of social dimensions is critical as this work makes initial strides in detecting minority-targeting online abuse. Continual reassessment of systems and their real-world influences remains essential even beyond research contexts. And usage policies must be crafted thoughtfully before any operational deployment. We believe impactful technology like this carries with it a profoundly ethical mandate. Progress ceases to be progress if attained through forfeiture of our values.

References

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. AL-BETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

Bharathi Raja Chakravarthi. 2023. Detection of Homophobia and Transphobia in YouTube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect Homophobia and Transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Saranya Rajiakodi, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Kishore Kumar Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on Homophobia and Transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, and Daniel García-Baena. Overview of the shared task on hope speech detection for Equality, Diversity, and Inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, and Nitesh Jindal. 2023. Overview of second shared task on Homophobia and Transphobia detection in English, Spanish, Hindi, Tamil, and Malayalam. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language*

Technology for Equality, Diversity and Inclusion, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022c. Overview of the shared task on Homophobia and Transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of Homophobia and Transophobia in multilingual YouTube comments. *arXiv preprint arXiv:2109.00227*.

Sarah Diefendorf and Tristan Bridges. 2020. On the enduring relationship between masculinity and homophobia. volume 23, pages 1264–1284.

Calogero Giametta and Shira Havkin. 2021. Mapping homo/transphobia: The valorization of the lgbt protection category in the refugee-granting system. *ACME: An International Journal for Critical Geographies*, 20(1):99–119.

Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anstasopoulos, and Marcos Zampieri. 2023a. OffMix-3L: A novel code-mixed dataset in Bangla-English-Hindi for Offensive Language Identification. *arXiv preprint arXiv:2310.18387*.

Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023b. nlpbdpatriots at blp-2023 task 2: A transfer learning approach to Bangla Sentiment Analysis. In *The First Workshop on Bangla Language Processing (BLP-2023)*, page 286.

Raviraj Joshi. 2022. L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus, Marathi BERT language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101, Marseille, France. European Language Resources Association.

Raviraj Joshi. 2023. L3cube-Hindbert and Devbert: Pretrained bert transformer models for devanagari based Hindi and Marathi languages.

Habibe Karayigit, Ali Akdagli, and Cigdem Inan Aci. 2022. Homophobic and hate speech detection using Multilingual-BERT model on Turkish social media. *Inf. Technol. Control.*, 51:356–375.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and Transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, page 100041.

Prasanna Kumar Kumaresan, Ruba Priyadharshini, Bharathi Raja Chakravarthi, Paul Buitelaar, Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Saranya Rajiakodi, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, Rafael Valencia-García, Kishore Kumar Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on Homophobia and Transphobia detection in social media comments". In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*. European Chapter of the Association for Computational Linguistics.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Hala Mulki and Bilal Ghanem. 2021. Let-Mi: An Arabic levantine Twitter dataset for Misogynistic language.

Nick Doiron. 2023. Hindi-BERT (revision c54eb83).

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

V Paul Poteat, Jillian R Scheer, Craig D DiGiovanni, and Ethan H Mereish. 2014. Short-term prospective effects of Homophobic victimization on the mental health of Heterosexual adolescents. *Journal of youth and adolescence*, 3.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023a. SentMix-3L: A Bangla-English-Hindi code-mixed dataset for Sentiment Analysis. *arXiv preprint arXiv:2310.18023*.

Md Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023b. nlpbdpatriots at blp-2023 task 1: A two-step classification for violence inciting text detection in Bangla. In *The First Workshop on Bangla Language Processing (BLP-2023)*, page 179.

Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023c. Offensive Language Identification in Transliterated and code-mixed Bangla. In *The First Workshop on Bangla Language Processing (BLP-2023)*, page 1.

Malavika Shetty. 2003. Language contact and the maintenance of the Tulu language in South India. In *Proceedings of the Eleventh Annual Symposium About Language and Society Austin (SALSA XI)*. Citeseer.

Juan Vásquez, Scott Andersen, Gemma Bel-Enguix, Helena Gómez-Adorno, and Sergio-Luis Ojeda-Trueba. 2023. Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214.

Antonio Ventriglio, João Mauricio Castaldelli-Maia, Julio Torales, Domenico De Berardis, and Dinesh Bhugra. 2021. Homophobia and mental health: a scourge of modern era. *Epidemiology and Psychiatric Sciences*, 30:e52.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sidney G-J Wong, Matthew Durward, Benjamin Adams, and Jonathan Dunn. 2023. cantnlp@ lt-edi@ ranlp-2023: Homophobia/transphobia detection in social media comments using spatio-temporally retrained language models. *arXiv preprint arXiv:2308.10370*.

A   Confusion Matrix
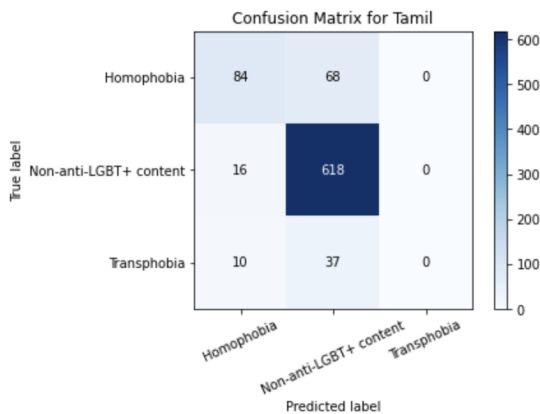
Figure 2: Confusion Matrix for Tamil Language
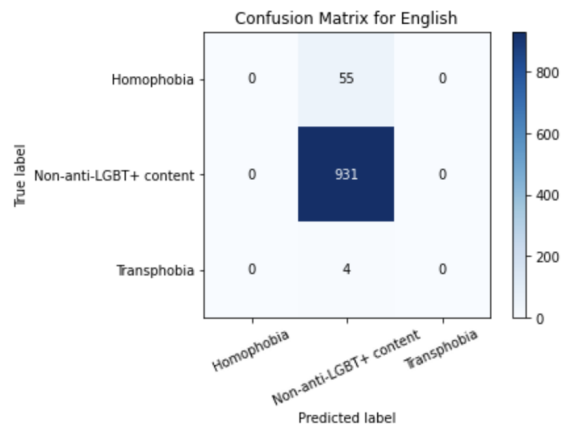
Figure 3: Confusion Matrix for English Language
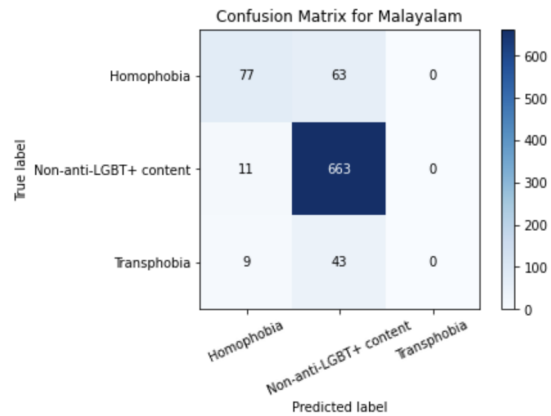
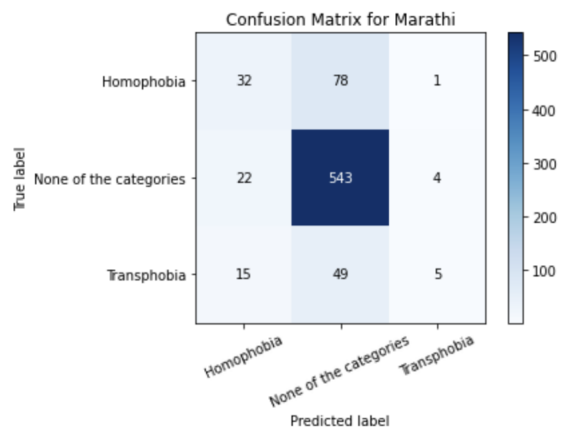Figure 4: Confusion Matrix for Malayalam Language

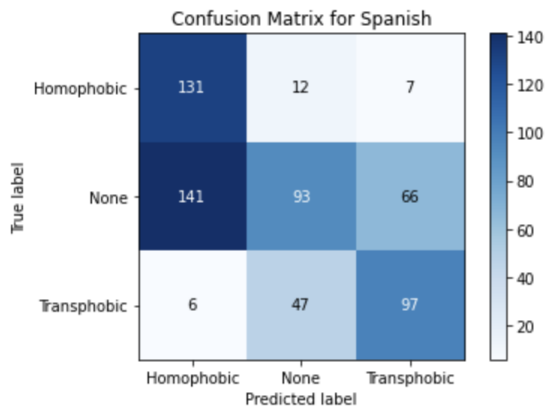Figure 5: Confusion Matrix for Marathi Language

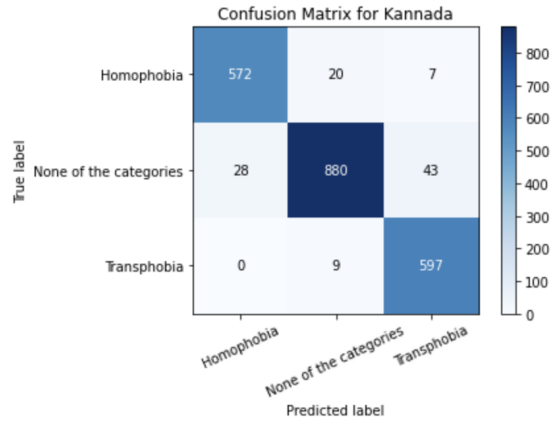Figure 6: Confusion Matrix for Spanish Language

Figure 9: Confusion Matrix for Kannada Language
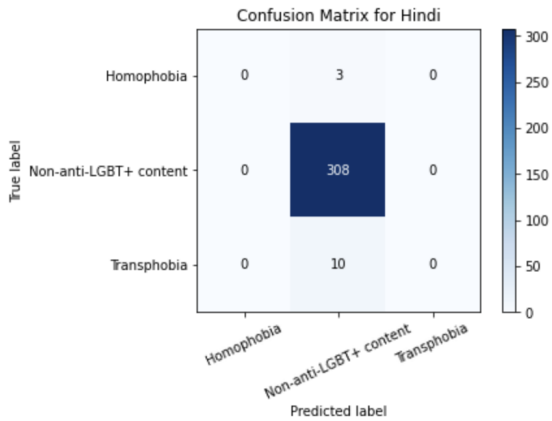

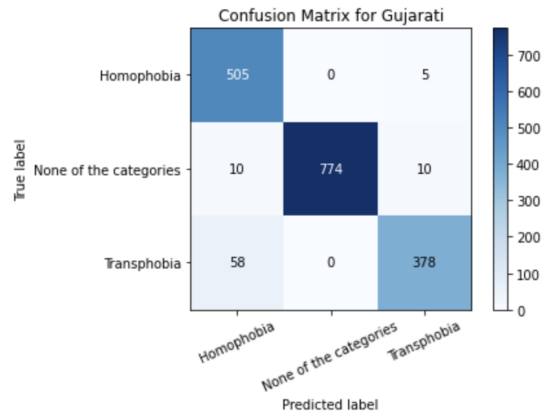Figure 7: Confusion Matrix for Hindi Language

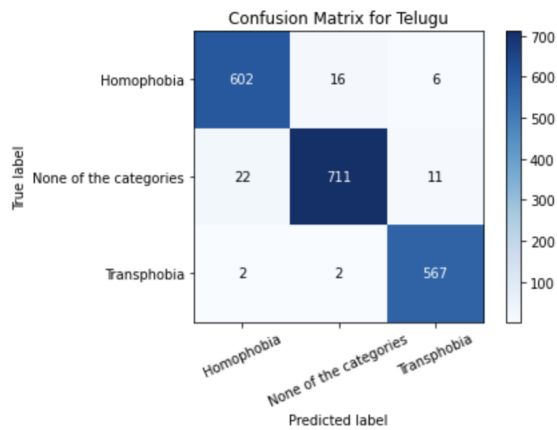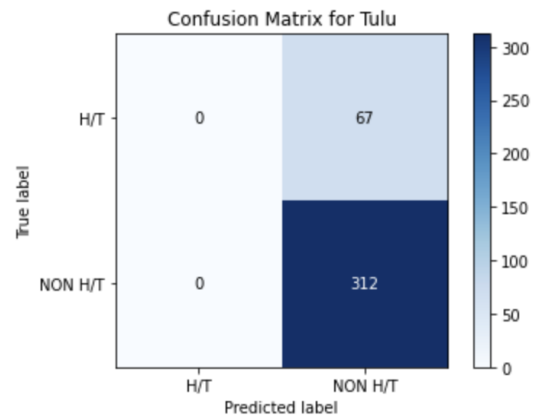Figure 10: Confusion Matrix for Gujarati Language

Figure 8: Confusion Matrix for Telugu Language

Figure 11: Confusion Matrix for Tulu Language